

Columbia Law School

## Scholarship Archive

---

Faculty Scholarship

Faculty Publications

---

2011

### The Measure of a MAC: A Quasi-Experimental Protocol for Tokenizing *Force Majeure* Clauses in M&A Agreements

Eric L. Talley

Columbia Law School, [etalley@law.columbia.edu](mailto:etalley@law.columbia.edu)

Drew O'Kane

[deokane@gmail.com](mailto:deokane@gmail.com)

Follow this and additional works at: [https://scholarship.law.columbia.edu/faculty\\_scholarship](https://scholarship.law.columbia.edu/faculty_scholarship)



Part of the [Banking and Finance Law Commons](#), [Business Organizations Law Commons](#), and the [Contracts Commons](#)

---

#### Recommended Citation

Eric L. Talley & Drew O'Kane, *The Measure of a MAC: A Quasi-Experimental Protocol for Tokenizing Force Majeure Clauses in M&A Agreements*, JOURNAL OF INSTITUTIONAL & THEORETICAL ECONOMICS, VOL. 168, P. 181, 2012; UC BERKELEY PUBLIC LAW RESEARCH PAPER NO. 1872568 (2011).

Available at: [https://scholarship.law.columbia.edu/faculty\\_scholarship/1699](https://scholarship.law.columbia.edu/faculty_scholarship/1699)

This Working Paper is brought to you for free and open access by the Faculty Publications at Scholarship Archive. It has been accepted for inclusion in Faculty Scholarship by an authorized administrator of Scholarship Archive. For more information, please contact [scholarshiparchive@law.columbia.edu](mailto:scholarshiparchive@law.columbia.edu).

# **The Measure of a MAC: A Quasi-Experimental Protocol for Tokenizing *Force Majeure* Clauses in M&A Agreements**

Eric Talley  
*UC Berkeley (Boalt Hall) School of Law*  
*etalley@law.berkeley.edu*

Drew O’Kane  
*UC Berkeley (Boalt Hall) School of Law*  
*deokane@gmail.com*

June 2011 (Version 1.1)

**Abstract:** We develop a protocol for using a well known lawyer-coded data set on Material Adverse Change/Effect clauses in acquisitions agreements to tokenize and calibrate a machine learning algorithm of textual analysis. Our protocol, built on both regular expression (RE) and latent semantic analysis (LSA) approaches, is designed to replicate, correct, and extend the reach of the hand-coded data. Our preliminary results indicate that both approaches perform well, though a hybridized approach improves predictive power even more. We employ Monte Carlo simulations show that our results generally carry over to out-of-sample predictions. We conclude that similar approaches could be used much more broadly in empirical legal scholarship, most specifically in the study of transactional documents in business law.

## 1. Introduction

In recent years, the field of empirical legal studies (ELS) has generated a wealth of academic scholarship that is impressive in both its breadth and reach. Though anomalous (at best) within the legal scholarship two decades ago, empirical methods now permeate virtually every area of law, including (but not limited to) administrative law, constitutional law, corporate / securities law, employment law, civil procedure, and jurisprudence.

To be sure, much of the reason for the recent growth and success of ELS is the influx of legally minded scholars with empirically sophisticated methodologies, including economists, psychologists, statisticians and sociologists. But in addition, empirical methods have also gained traction because of the expansion of publicly available data. As published opinions, regulations, transactional forms and other legal documents have become increasingly available in digital form in the mid 1990s, numerous scholars interested in the empirical dimensions of law had – for the first time – a significant amount of usable raw material to synthesize, interpret, parse, describe, analyze, and test.

And yet, a significant roadblock continues to hamper the success of empirical methods – the fact that empirical legal scholarship tends (by definition) to be “hard,” numerical and quantitative, while most original legal sources tend – like the practice of law itself – to be nuanced, textual and qualitative. Consequently, for those interested in pursuing empirical questions in law, it is still the norm to rely on human filters as transcription vehicles, asking students, researchers, or practicing attorneys to read, summarize, classify, and parse quantitative data from their original qualitative texts. In many ways, this human element is unavoidable (and even desirable), since the practice of law is in many ways the art of navigating between nuanced forms of expression and hard legal outcomes or predictions. Similarly, the process of hand coding allows those inputting data to make nuanced judgments about subtle differences in detail between raw data sources (such as transactional documents from different jurisdictions) and how those sources are coded into a quantitative dataset (e.g., are the applicable rules immutable in

one jurisdiction and default rules the other?). In most areas of law practice (and particularly technical fields), this appreciation for nuance is invaluable.

Yet exclusive reliance on human interpretation and hand-coding of legal texts can significantly impair one's ability to marshal the power of textual legal datasets. Perhaps the first of these roadblocks is cost. Unlike other forms of hand-coding, human coding of legal sources generally requires personnel with legal expertise in the practical / jurisdictional area of interest. Within most developed legal systems, this generally means enlisting practicing lawyers or advanced law students to do the work. But for either group, opportunity costs and outside prospects can be significant, driving up the costs of collection. Moreover, such datasets not only entail significant startup expenditures, but their marginal maintenance and updating costs remain high as well. Hand coding raw legal/regulatory data is an arduous and painstaking process, often taking weeks or months to code a relatively modest set of variables.

A related roadblock concerns internal reliability and consistency of a hand-coded dataset: given the human capital and time requirements of hand-coding, large-scale, longitudinal projects, such endeavors often involve a revolving door of personnel. The practice of making judgment calls, relative attention to nuance, and the legal backgrounds of contributors vary wildly, both between coders and over time. It is often difficult if not impossible to know the nature / direction of resulting biases, rendering corrective measures challenging at best, prohibitive at worst.

During much of the time in which empirical legal scholarship has been developing, computer scientists and natural language theorists have been developing tools (largely – though not exclusively – outside of legal applications) for the large-scale automated analysis of textual data. Pioneering developments in these fields, many of which have emanated during the last fifteen years, have already come to dominate much of computational biology, library sciences, and information technology. (Indeed, the core business of companies like Google has and remains in the management, organization and indexing of vast quantities of qualitative data). Although natural language processing approaches are now starting to infiltrate legal practice and scholarship (particularly in the areas of e-discovery and analysis of judicial opinions), they have tended to do so in a way

that is independent of and parallel to traditional forms of database development within legal scholarship. In most private / transactional law contexts, moreover, these approaches are virtually nonexistent.

In this paper, we propose a quasi-experimental method for using expensive, lawyer-coded databases to calibrate (or “tokenize”) a machine learning protocol for replicating, correcting, and significantly expanding those databases. Effectively, our approach treats existing hand-coded data as a type of “field experiment,” embodying assessments, nuanced judgments and resulting regularities that can inform the basis of a set of logical conventions that facilitate automated coding, with potentially greater reliability, lower cost, greater scalability, and enhanced speed. Our specific focus is on mergers and acquisitions (M&A) agreements executed between 2007 and 2008, and in particular the use of (so-called) “Material Adverse Change” or “Material Adverse Event” (MAC/MAE) provisions in such agreements. MAC/MAEs are a central workhorse provision for allocating risk and uncertainty that can emerge between the execution of a corporate acquisition and its closing – a period that can often take many months (if not years). Such provisions typically condition one party’s (almost always the buyer’s) obligation to close the transaction on the absence of any occurrence, condition, change, event or effect that materially and adversely affects some enumerated dimension of the deal’s value. When triggered, MAC/MAE provision effectively gives the advantaged party the right to “walk away” from a deal (or at least to threaten to do so as a backdrop to renegotiation). Consequently, MAC/MAE provisions tend to be a central focus of negotiating parties during an acquisition – they are truly “dickered” as part of the deal. (See, e.g., Talley (2009); Gilson & Schwartz (2001); Macias (2009).

As a tokenizing/experimental dataset, we make use of a well known attorney-coded database that tracks the presence / absence of twenty different MAC/MAE sub-provisions in announced agreements executed between 2007 and 2008, and previously analyzed in Talley (2009). Combining this data with the raw text from each deal’s MAC/MAE provision, we develop two predictive machine learning protocols – both built largely on Python computer code – for diagnosing the presence of each sub-provision. The first protocol is based on a “Regular Expression” (RE) algorithm – a Boolean

dictionary that summarizes syntactical patterns that are characteristic of each type of contractual sub-provision. The second protocol utilizes “Latent Semantic Analysis” (LSA) techniques for analyzing the raw content of each provision by generating a frequency table of its terms (a metaphorical “bag of words” inventorying the terms used in each MAC/MAE provision). Each of the RE and LSA approaches is capable of generating relatively accurate calibrations that perform reasonably well in replicating the hand-coded data (both achieved overall classification accuracies of approximately 80%). However, we further demonstrate that when used in combination with one another, RE and LSA methodologies perform even better (producing classification accuracies in the 85% range). We also employ Monte Carlo methods to simulate the out-of-sample predictive power of each methodology (and both in combination), and our results remain qualitatively similar (though, not surprisingly, a bit weaker and noisier than in the full-sample calibration). Overall, on the basis of this exercise, we conclude that our quasi-experimental protocol provides a promising approach for replicating, correcting, and vastly expanding the depth and breadth of existing hand-coded legal datasets, at a significant marginal cost savings.

Before proceeding, an important caveat deserves explicit attention. Our enterprise in this paper employs a methodology that (we think) is appropriately classified as “experimental” in nature; but by our own admission, this is far from a conventional use of the term. We effectively treat an existing hand-coded database as a field experiment, which we then utilize to calibrate and tokenize a predictive model, which in turn is capable of extending the initial data’s scope and breadth, even beyond their initial boundaries. It is clear that our study design parts company with conventional experimental protocols, particularly in relaxing the element of control over the experimental setting and manipulations thereof (Camerer & Talley 2008). Consequently, our approach cannot make large contributions for questions where experimental approaches are thought to be most apt: the task of developing and/or testing causal theories of behavior. On the other hand, our enterprise here is not (or at least is not yet) about testing causal theories. Rather, we endeavor to use quasi-experimental

data to calibrate a diagnostic and predictive model of contractual provisions in real world contracts.

Moreover, the general approach championed here can be easily carried over to other experimental settings. Indeed, nothing requires the tokenizing database to be drawn from an existing data source – it could also come from more conventional experimental manipulations (both in and outside the laboratory). For example, we are currently piloting an extension to this study where subjects in a laboratory are asked to evaluate, negotiate, and “price out” a set of specimen MAC/MAE provisions based upon our dataset within a hypothetical deal. This approach – if successful – will allow us not only to calibrate a predictive model of the presence / absence of certain canonical provisions (as done here), but it will permit us to assemble a cardinal monetary measure of the “buyer friendliness” of each provision based on its constituent parts -- one that can possibly also extrapolated outside the experimental sample. We therefore view the current project as a first (but important) step in combining experimental data with machine learning approaches for large quantitative text analysis in commercial and corporate law.

Our analysis proceeds as follows. Section 2 describes MAC/MAE provisions in greater detail, and briefly discusses their significance in corporate law and M&A practice. Section 3 describes our data and general methodology, including the RE and LSA protocols we develop for building our algorithmic models. Section 4 presents our (tentative) results, for both full-sample calibrations and for simulated out-of-sample predictions using Monte Carlo methods. Section 5 discusses our results in greater detail, considers a number of extensions to our approach, and concludes.

## **2. Background: What *Is* a MAC?**

As noted above, our paper focuses on using a pre-existing lawyer-coded database of MAC/MAE provisions as a quasi-experimental template for tokenizing natural language processing protocols. To better motivate our enterprise, this section spends

some time describing what, exactly, these provisions are, what purposes they serve, and how they come into existence.

MAC/MAEs are a species of contractual *force majeure* (or “act of God”) provision that are peculiar to acquisitions and financing transactions. Put simply, the MAC/MAE functions as a type of express condition on one party’s (or sometimes both parties’) obligation to complete performance an already-executed contract. As a matter of legal formality, most MAC/MAEs in M&A and financing deals are construed to be *conditions subsequent*: i.e., the occurrence of the enumerated contingency relieves the advantaged party of her pre-existing duty to close. In some situations, however, the MAC/MAE can be construed as a *condition precedent*, whereby the advantaged party has no duty to perform *unless* the enumerated contingency obtains. Although this difference may seem to be merely semantic, it has practical implications – for it effectively flips the evidentiary burden of demonstrating whether the condition has occurred from the advantaged party (in a condition subsequent) to the disadvantaged party (in a condition precedent).

Regardless of how it is construed for evidentiary purposes, the conventional architecture of a MAC/MAE provision is perhaps best described as resembling a (metaphorical) piece of Swiss cheese.<sup>1</sup> One portion of the provision (“affirmative” section) usually appears at the beginning of the provision, and constitutes the cheesy bit, enumerating often broad categories of contingencies where a material change in circumstance relieves the buyer of her obligation to close. Another portion of the MAC/MAE provision (the “exclusion” or “carve-out” section) invariably follows, and constitutes the metaphorical holes in the cheese. The exclusions typically take the form of a more lengthy and specific list of enumerated contingencies that *do not* constitute an escape hatch for performance, notwithstanding the broad affirmative provisions.

The MAC/MAE provision in the 2007 acquisition of the Huntsman Corporation (which culminated in the litigated case of *Hexion v. Huntsman* (2008)) provides an apt example of this architecture. The MAC/MAE from that deal is reproduced in its entirety

---

<sup>1</sup>We give due acknowledgment co-author Talley’s 9-year old daughter, Gracie, whose culinary obsession with Mac and Swiss cheese provided the initial inspiration for this metaphor.



below, with the affirmative portion of the highlighted in blue (dark grey), and the carve-outs in red (light grey):

A “Company Material Adverse Effect” means any occurrence, condition, change, event or effect that is materially adverse to the financial condition, business, or results of operations of the Company and its Subsidiaries, taken as a whole; provided, however, that in no event shall any of the following constitute a Company Material Adverse Effect: (A) any occurrence, condition, change, event or effect resulting from or relating to changes in general economic or financial market conditions, except in the event, and only to the extent, that such occurrence, condition, change, event or effect has had a disproportionate effect on the Company and its Subsidiaries, taken as a whole, as compared to other Persons engaged in the chemical industry; (B) any occurrence, condition, change, event or effect that affects the chemical industry generally (including changes in commodity prices, general market prices and regulatory changes affecting the chemical industry generally) except in the event, and only to the extent, that such occurrence, condition, change, event or effect has had a disproportionate effect on the Company and its Subsidiaries, taken as a whole, as compared to other Persons engaged in the chemical industry; (C) the outbreak or escalation of hostilities involving the United States, the declaration by the United States of war or the occurrence of any natural disasters and acts of terrorism, except in the event, and only to the extent, of any damage or destruction to or loss of the Company's or its Subsidiaries' physical properties; (D) any occurrence, condition, change, event or effect resulting from or relating to the announcement or pendency of the Transactions (provided, however, that this clause (D) shall not diminish the effect of, and shall be disregarded for purposes of, the representations and warranties relating to required consents, approvals, change in control provisions or similar rights of acceleration, termination, modification or waiver based upon the entering into of this Agreement or consummation of the Merger); (E) any change in GAAP, or in the interpretation thereof, as imposed upon the Company, its Subsidiaries or their respective businesses or any change in law, or in the interpretation thereof; (F) any occurrence, condition, change, event or effect resulting from compliance by the Company and its Subsidiaries with the terms of this Agreement and each other agreement to be executed and delivered in connection herewith and therewith (collectively, the "Transaction Agreements"), actions permitted by this Agreement (or otherwise consented to by Parent) or effectuating the Financing; or (G) any occurrence, condition, change, event or effect resulting from or in connection with any Divestiture Action...

Even a cursory inspection of this provision yields a few immediate observations. First, while this provision is of roughly average length (clocking in at 422 words), it is far from Hemingwayesque, and it contains significant detail. Second, the affirmative provision is brief and drafted in sweeping terms (applying to a shock in circumstances that materially affects the seller's financial condition, business, or results of operations); the carve-outs / exceptions, in contrast are spelled out in great, precise, tedious details. Although this MAC/MAE provision was recognized even at the time as having a fairly large number of seller friendly carve-outs,<sup>2</sup> this general pattern persists across all deals studied here.

---

<sup>2</sup>N.Y. Times DealBook, “Huntsman-Hexion: A Deal Agreement to Applaud,” available for download at <http://dealbook.blogs.nytimes.com/2008/01/11/huntsman-hexion-a-deal-agreement-to-applaud/> (Jan. 11, 2008, 16:34 EST).

Within the larger architecture of a merger agreement, a MAC/MAE might be found in one of many different locations. In some cases it can be found as a stand-alone provision, delineated separately from other terms, and specifically granting the favored party a contingent right to walk away. In other deals, the MAC/MAE is incorporated into the representations and warranties, explicitly tied to a “bring down” provision that effectively scuttles the merger when the MAC/MAE is triggered. In yet other cases, the MAC/MAE appears as an embedded component of the closing conditions to a deal. In yet other deals, the MAC/MAE may be found spread across two or more of these sections of an acquisition agreement.

As the Huntsman excerpt implicitly suggests, negotiating teams often spend a significant amount of time dickering the precise terms of MAC/MAE provisions. The provisions have accordingly garnered a great deal of attention from both academics and legal professionals. Theories abound as to what purpose the MAC/MAE plays in an acquisitions agreement that is unique from other risk allocation devices (such as contingent prices, earn outs, termination fees, indemnities and guarantees, and the like). One prominent theory (Gilson & Schwartz 2005) posits that MAC/MAE provisions optimally impose risk on the target’s management, which would otherwise have poor incentives to maintain firm value in the interim period between execution and closing. Others (Talley 2009) have argued that MAC/MAE provisions are uniquely well suited to allocate *ambiguity* about the mutual gains from the deal (as distinct from risk). Still others (Choi & Triantis 2009) have argued that MAC/MAE provisions are calibrated to facilitate deal restructuring with minimal transaction costs. It is likely that each of these explanations plays a role in explaining the purposes behind MAC/MAE provisions. For the purposes of this paper, we need not adjudicate among these competing theories. We simply note that each of them is, in principle, testable with sufficient empirical data. And the purpose of this project is to suggest a way to enhance and improve that data.

### 3. Data and Methodology

The significance that MAC/MAE provisions have for practicing lawyers is reflected in the considerable interest within private and academic circles. Notable among these efforts is a longstanding database built by the New York law firm of Nixon Peabody LLP, a firm with a substantial mergers and acquisitions practice. In part as a client development service, Nixon Peabody produces an annual survey of MAC provisions, publishing summary statistics in an annual publication used widely in the industry.<sup>3</sup> The firm has been producing its annual survey for over a decade<sup>4</sup>, and since 2005 its methodology has become sufficiently consistent to be somewhat usable in time series analysis. We were given access to this database, and sampled 123 acquisitions agreements coded in the 2007-08, corresponding with those involving public targets and for which we could obtain the full merger agreements from publicly available sources (usually the SEC’s Edgar database). The plurality of the deals (45.5%) involve stock mergers, followed by cash mergers (30%) negotiated tender offers (11%), stock purchases (9%) and asset sales (4%). This appears representative of deals occurring during the same time span (Talley 2009). In what follows, we will periodically refer to the Nixon-Peabody lawyer coded data set as our “experimental” database.

All coding for the Nixon Peabody data was done by practicing members of the New York bar, usually (but not wholly) comprised of mid-level associates.<sup>5</sup> For each MAC/MAE in a deal  $i$ , the coding attorney was required to identify a *vector*  $\{y_{i1}, y_{i2}, \dots, y_{ik}\}$  of binary “attributes”, effectively indicating the presence / absence of specifically enumerated sub-provisions that might be included in the MAC/MAE. Consistent with the architecture described above, attorneys were asked to code both “affirmative” MAC/MAE clauses (e.g., terms that deal with the target’s financial condition, the seller’s ability to close the deal, the target company’s prospects, etc.), and exceptions or “carve

---

<sup>3</sup> Nixon-Peabody (2008), [http://www.nixonpeabody.com/publications\\_detail3.asp?ID=2488](http://www.nixonpeabody.com/publications_detail3.asp?ID=2488).

<sup>4</sup>See *id.* at 2.

<sup>5</sup>Unfortunately, Nixon-Peabody does not keep track of individual coding attorneys, and we are therefore unable to control for any coder-specific biases. That said, we performed numerous hand audits of this data in order to ensure its consistency as much as possible (see discussion below).

outs” (such as changes to the economy in general, changes to securities markets, interest rates, GAAP, etc). As a general matter, the presence of affirmative MAC/MAEs tends to expand the breadth (and the buyer-friendliness) of the provision, while carve outs tend to contract that breadth. For our analysis, we focused on 20 coded provisions, selected (largely) at random from the entire set of 44 MAC/MAE provisions coded in the Nixon-Peabody dataset. An inventory, brief descriptions and sample frequencies of these provisions are contained in Table 1 below.

[INSERT TABLE 1 HERE]

As Table 1 illustrates, there are seven affirmative MAC/MAE provisions and thirteen carve-outs included in our experimental sample. They range considerably in frequency across deals, from a high of 94.3% (for the affirmative MAC on business, operations, and financial condition -- “MBOF”) to less than 3% (for affirmative MACs on target losses over a specified threshold -- “MExcessLos”, and a MAC on the target’s business prospects -- “MPrspects”). Moreover, there appears to be considerable variation across categories, as indicated by the correlation matrix across provisions in Table 2, below. Most provisions have weak positive correlation with others, and the mean pair-wise correlation across terms is approximately 0.25.

[INSERT TABLE 2 HERE]

Our approach utilizes (and ultimately combines) two distinct methodological protocols for classifying textual data: (1) Regular expressions (RE) or “Boolean” protocols; and (2) Latent Semantic Analysis (LSA). We describe each of them below, *ad seriatim*. The RE protocol we developed is based upon a hand-built “dictionary” of typical/characteristic grammatical patterns that frequently appear across texts, and which constitute alternative ways to insert a particular type of provision. Practically speaking, this dictionary very much has the look and feel of a compilation of Lexis / Westlaw queries that are applied to the sample MAC/MAE documents for each coded term of

interest. For example, in approximately 30% of the experimental data, the MAC/MAE contains a carve-out provision for “Acts of God” (see Table 1). We found we were able to identify approximately three-quarters of these instances with a simple search protocol that tests for the proximity (e.g., within five words) of the word *God* and various conjugations of the word *act*. Other provisions, in contrast, exhibit greater linguistic and grammatical heterogeneity, and require more elaborate – and often more numerous – conditional search protocols. Thus, a substantial portion of our Python code consists of functions that invoke customized Boolean dictionaries to diagnose the presence or absence of each type of lawyer-coded provision in each MAC/MAE. (As one might surmise, the process of building these dictionaries proved painstaking, involving nuanced reading of example provisions and common sense. We are still working on ways to further refine these dictionaries.) For notational housekeeping, consider provision  $i$  in our data which with lawyer-coded attributes  $\{y_{i1}, y_{i2}, \dots, y_{iK}\}$ . Our regular expression protocol similarly generates a vector of binary diagnostic predictions about the presence of each sub-provisions of the MAC. We denote these predictions with  $\{r_{i1}, r_{i2}, \dots, r_{iK}\}$ .

The second protocol was much less structured, and drew on machine learning techniques from latent semantic analysis (LSA) literature. In contrast to RE, the LSA approach pays little heed to linguistic architecture of a document (e.g., grammar and usage), and concentrates instead on the “raw materials” (e.g., words) that compromise that document. To implement our LSA approach, our Python code extracted, for each MAC/MAE provision, a unigram frequency inventory – literally a “bag of words” concatenation index, tabulating counts of each unique word across the entire set of deal documents. For the entire dataset, this process resulted in a term frequency matrix tracking the raw word counts of approximately 3,000 unique unigrams used across all documents.<sup>6</sup> Denote this matrix by  $\mathbf{N}$ , where representative element  $n_{ij}$  represents the number of times term  $j$  appears in document  $i$ .

---

<sup>6</sup> In an unreported robustness check, we generated (and transformed) raw count not only for single-term unigrams, but also for bigrams and trigrams of consecutive terms. This alteration substantially increased computing time (generating over 30,000 word frequency variables), while only marginally enhancing the predictive power of our model. We therefore confine our analysis below to the case of single term unigram frequencies.

Next, and following conventional practice, we transformed the elements of  $\mathbf{N}$  from their raw frequency counts into “term frequency - inverse document frequency” (or TF-IDF) measures. The resulting transformed matrix,  $\mathbf{T}$ , contains representative element  $t_{ij}$  for document  $i$  and term  $j$ , defined by

$$t_{ij} = \left( \frac{n_{ij}}{\sum_m n_{mj}} \right) \times \ln \left[ \frac{|\{j:n_{ij}>0\}|}{M} \right]^{-1}, \quad (1)$$

where  $m \in \{1, \dots, M\}$  indexes the universe of documents analyzed. The purpose and effect of this transformation is to accord greater proportional weight to terms that appear with large frequency in a particular document *and yet* are relatively uncommon overall. The first bracketed element of (1) represents the raw count of a given term in document  $i$  relative to its total across all documents. The second term consists of the log of the inverse frequency with which term  $j$  appears (at least once) across the universe (with cardinality  $M$ ) of documents being analyzed. By “rewarding” the frequent intra-document use of terms that are rare on the whole, the TFIDF transformation tends to be better able to differentiate unique documents (Salton & Buckley 1988).<sup>7</sup>

Because the TF-IDF transformation in (1) has a fixed point at  $n_{ij}=0$ , the transformed matrix  $\mathbf{T}$  remains both extremely large and extremely sparse. Following conventional approaches in LSA, we therefore proceeded to reduce the dimensionality of  $\mathbf{T}$  through singular value decomposition – a generalized form of principal component analysis. We retained the factors corresponding to the largest six Eigenvalues from the decomposition. Ultimately, for each MAC/MAE provision  $i$ , our singular value decomposition of  $\mathbf{T}$  resulted in the generation of factor matrix  $\mathbf{X}$  with representative row  $\mathbf{x}_i$  consisting of the 6-tuple  $\{x_{1i}, x_{2i}, \dots, x_{6i}\}$ .<sup>8</sup>

---

<sup>7</sup> We found little difference in results regardless of whether we utilized the matrix of raw unigram counts  $\mathbf{N}$  or the TFIDF transformed matrix  $\mathbf{T}$ . Nevertheless, because this transformation is routinely applied in the natural language processing literature, we employ it in the analysis that follows.

<sup>8</sup> Note that the factors  $\mathbf{x}_i$  emerging from the singular value decomposition have no natural interpretive content, as they are merely algebraic artifacts of the underlying composition of  $\mathbf{T}$ . Accordingly, we will spend no time exploring the intuition behind the estimated coefficients on  $\mathbf{x}_i$ .

We are now in a position to use the RE and/or LSA protocols to predict the presence of a specific term in the experimental (i.e., lawyer coded) database. As noted above, one interesting advantage of using both protocols is that it becomes possible to marshal their combined explanatory power in a hybrid predictive model. Thus, for each MAC/MAE provision  $k$ , the baseline empirical specification for all of our predictive regressions is one that allows such combination, and is as follows:

$$\Pr\{y_{ik} = 1\} = f(\alpha + \beta \cdot r_{ik} + \gamma \cdot \mathbf{x}_i), \quad (2)$$

where  $y_{ik}$ ,  $r_{ik}$  and  $\mathbf{x}_i$  are as described above, and  $\alpha$ ,  $\beta$  and  $\gamma$  are coefficients. All results reported below utilize a logistic maximum likelihood model (though our results appear similar under probit and linear probability models as well).

## 4. Tentative Results

### a. Full Sample Calibration

Consider first the estimation of (2) across our entire dataset. This approach generates the most complete calibration measure of our machine learning protocols to the experimental data set. (It does not, however, allow us for test for out-of-sample predictions, however – a question we turn to below). Table 3 reports on the prediction characteristics of our logit estimation of equation (2). The table suppresses the direct logit estimation results, as the estimated coefficients are of little interest and most cannot be easily interpreted. Rather, it illustrates predictive performance across all 20 sampled MAC/MAE terms, and three different specifications of (2). (Thus, each row / panel entry in Table 3 represents a separate estimation). In the first panel, we drop all the LSA variables, and regress the lawyer-coded attributes solely on our RE predictor. In the second panel, we drop the RE diagnostic and regress solely on the LSA factors. Finally, in the third panel, we include both the RE and the LSA factors in a “hybrid” model.

We consider two measures to evaluate predictive performance. First, we measure “correct” categorization rates using an assignment protocol that predicts the presence of a term in the experimental data if its predicted probability (at the estimated coefficients) exceeds  $\frac{1}{2}$ . Second, we compute the Receiver Operating Characteristic (ROC) curve, and derive the area under the curve (AUC). (The ROC is a graphical plot of the false positive classification rate against the true positive classification rate as one continuously varies the critical probability threshold for prediction assignment from 0.0 to 1.0. A model that predicts nearly perfectly will exhibit a highly concave ROC curve, with an area under the curve of close to 1. A model that makes nearly random predictions will exhibit a ROC that is approximately linear, with an area under the curve of close to 0.5).<sup>9</sup>

[INSERT TABLE 3 HERE]

Our estimates produce a respectable (though still imperfect) rate of correct classification in replicating the Nixon-Peabody data set across the 123 sampled deals. Overall, each of the RE and LSA approaches were able to match the sample terms with an average 80 percent accuracy rate. On a term-by-term basis, our mismatch rate range between zero and thirty-seven percent. (Not surprisingly, misclassifications in the RE approach are skewed towards false negatives, while they are generally balanced between false negatives and false positives in the LSA estimations). On the whole, the LSA estimates tended to yield slightly larger correct classification rates than the RE estimates, though in a few situations the LSA predictors preformed much worse. When the RE and LSA approaches were combined, however, classification accuracy generally increased (to 84%), and in some cases the improvement was dramatic.

The ROC-AUC measures suggest a similar pattern. In all our specifications, estimated ROC areas indicated that each of our approaches is diagnostically probative. Moreover, the combination of RE and LSA approaches delivers a discernible increase in predictive performance across nearly all coded terms. We view these base results as a

---

<sup>9</sup> As Hanczar et al (2010) recently demonstrated, the ROC-AUC metric can suffer from being relatively noisy in smaller data sets. We nevertheless use it here as a general guidepost, especially in light of the absence of alternative good measures of predictive probity.



promising start, but one that can be significantly improved upon with more consultative and programming attention. In fact, there may be a sense in which these figures *understate* the accuracy of our approach, since the Nixon-Peabody data invariably will contain undetected coding errors. Auditing some of the evident mismatches, we discovered that Nixon-Peabody attorneys appear to have had a mis-coding rate in excess of 3%. Consequently, it is plausible that Table 3 understates the correct categorization rate relative to the “true” underlying contractual terms.<sup>10</sup>

#### b. Out of Sample Monte Carlo Prediction

Although Table 3 illustrates the explanatory power gained by marshaling both regular expression and latent semantic techniques for predicting the presence / absence of particular terms, its results are distinctly *within sample*. They need not (and likely do not) remain as strong when the predictive model is taken outside of the sample constellation of deals. Yet it is predominantly in out-of-sample prediction where our approach can be useful in economizing the time and energy consumed by hand coding. The discussion below, therefore, considers out-of-sample prediction issues more squarely.

In order to simulate out of sample prediction, we employed a Monte Carlo bootstrap aggregation (“bagging”) approach proposed originally by Brieman (1996). We devised an identical bagging protocol for each contractual provision studied (i.e., each of the 20 affirmative MACs/MAEs and exceptions). Within each Monte Carlo iteration, the data were randomly segregated into two groups: A “training” dataset, consisting of roughly 75 percent of our observations, and a “testing” dataset, consisting of the remaining 25 percent of the data. We then fit equation (2) to the “training” data using (successively) LSA and hybrid approaches. And as before, we used the coefficient estimates to make probabilistic predictions of the presence/absence of the contractual term at issue in the remaining testing data, generating estimates of both correct classification rates, and of ROC-AUCs. For each of the 20 terms studied, we repeated the

---

<sup>10</sup> This assertion, of course, requires significantly more investigation. For example, we have not yet audited any of the evidently correctly matched terms to determine whether *both* the Nixon-Peabody data *as well as* our own protocol are mis-coding some deals.

Monte Carlo simulation 1,000 times.<sup>11</sup> (Each succeeding iteration re-sampled our training/testing data with replacement.) Table 4, therefore, reports on the resulting empirical distributions of both classification rates and ROC-AUCs. (Note that the table reports only on the “pure” LSA and the hybrid model, excluding the pure RE model – given the way that the RE dictionary was assembled, Monte Carlo methods were not informative for the pure RE model).

[INSERT TABLE 4 HERE]

As expected, the out of sample predictions in Table 4 are weaker and noisier than the full-sample calibrations of Table 4. Moreover, for two provisions (MExcessLoss and MProspects), there was simply not enough data variability within the sample to execute the bagging protocol with a training dataset with any consistency. Nevertheless, for the remaining terms, our simulated out of sample predictions appear to remain relatively strong. Average correct classification rates across all terms are approximately 73.5% for the LSA specification and 78% for the hybrid specification. The ROC-AUC measures also appear relatively good – only slightly smaller numerically than those in Table 3, but subject to considerably more noise.

In unreported robustness checks, we reran the bagging protocol with varying proportions of training / testing data. The results were largely consistent – though they tended to weaken as our training database shrank below 50% of the entire sample. We conjecture that as we add additional MAC/MAE provisions to our database, we will be able to enhance this performance significantly.

## **5. Discussion, Extensions and Conclusion**

Although we consider the above exercise to be a successful first step, it is limited by a number of factors – not the least of which is sample size. With only 123 coded deals

---

<sup>11</sup> Based on preliminary investigations, very little additional accuracy emerged from increasing the number of iterations to either 10,000 or 100,000.

to work with, we were unable to marshal much of the potential power of both RE and LSA protocols (particularly the latter) in calibrating our predictive models. We are currently working to expand the size of our experimental data over time and cross-sectionally.

Similarly, we have limited our attention here to deals that are both publicly available and coded in the Nixon-Peabody dataset. An obvious follow-on step would be to take our predictive model outside this data set, applying it to *all* publicly available merger agreements (be they included in the experimental data or not). We are currently in the process of implementing this step for the years 2007-10, and will report on results in a later draft of this paper.

A third extension of our approach would be to utilize alternative existing data to provide an quasi-experimental calibration device for tokenizing a larger database. Although the Nixon-Peabody data is rich, detailed, and coded by practicing lawyers, it is not unique. In particular the American Bar Association also harvests a regular survey of M&A agreements (including MAC/MAE clauses) that could provide either a useful calibration check, or additional coding attributes not available in the Nixon-Peabody data.

Yet another factor that constrained our analysis concerned the inherent limitations in using an experimental data set (Nixon-Peabody, ABA or something else) that is not collected under conventional experimental conditions. As noted above, we were unable to exercise control over either the conditions under which lawyers coded this data, or the targets of their efforts. Thus, we were unable to track subject-specific effects or other systemic factors that may have affected the reliability of the coding efforts. Moreover, we could not glean other – potentially more interesting – metrics for the breadth and content of a MAC/MAE provision. For instance, the Nixon-Peabody data does not elicit an all-things-considered assessment of whether a particular provision is buyer- or seller-friendly, or how complex and unpredictable its application might be in practice. Nor, for that matter, does the data elicit information about monetary tradeoffs that a representative buyer / seller might make in negotiating over a MAC/MAE provision (e.g., how much of a discount would a seller demand or a buyer accept in order to move from a relatively stingy MAC/MAE to one that gave the buyer additional outs?). We are currently piloting

a more controlled experimental instrument that attempts to elicit this information, and which will constitute a separate paper from this one.

Notwithstanding its significant recent growth, data-driven empirical methods in legal scholarship have only begun to scratch the surface of their ultimate capabilities. Litigated cases, appellate opinions, and Supreme Court decisions – the chief targets of much of the current ELS literature – are but a tiny fraction of what attorneys actually do in practice. Much transactional work, particularly within business law, is only now beginning to lend itself to serious quantitative analysis. This project presents some initial steps in pushing those efforts forward in the M&A context, and in a way that facilitates the harvesting of data at a lower cost, with greater consistency, and more dynamic adaptability than is currently the available. Based on our results thus far, we are optimistic that we can implement our protocol on a wide scale basis, not only to understand the nature and evolution of MAC / MAE clauses (an important topic itself), but also to facilitate the harvesting of data across a large array of legal transactional domains.

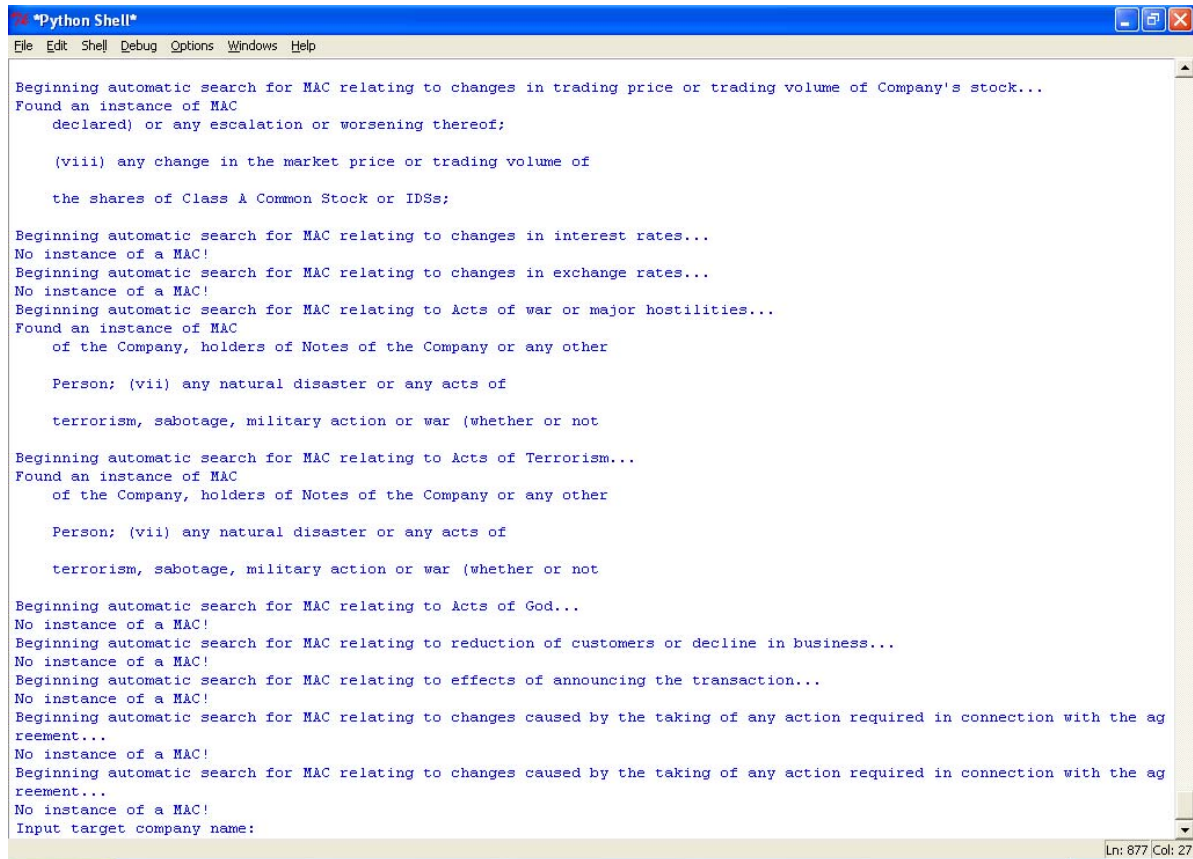
## References

1. Bishop, C.M., 2006. *Pattern Recognition and Machine Learning*. Springer.
2. Bishop, C.M. 1995. *Neural Networks for Pattern Recognition*. Oxford University Press.
3. Breiman, L., 1996. "Bagging Predictors." *Machine Learning* 24: 123-140.
4. Breiman, L., 2001. "Random Forests." *Machine Learning* 45: 5-32.
5. Burges, C.J.C., 1998. "A Tutorial on support vector machines for machine learning". *Knowledge Discovery and Data Mining* 2: 121-167.
6. Camerer C. and Talley E. (2007), Experimental study of law, Chapter 21 in: *Handbook of Law and Economics* (volume 2), A. Polinsky & S. Shavell, eds.
7. Cristianini, N. & J. Shawe-Taylor, 2000. *Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press.
8. Cybenko, G., 1989. "Approximation by Superpositions of a Sigmoidal Function." *Mathematics of Control, Signals and Systems*. 2: 303-314.
9. Freund, Y. & R. Schapire, 1996. "Experiments with a new boosting algorithm," in *13<sup>th</sup> International Conference on Machine Learning* pp. 148-156 (L. Saitta ed.).
10. Friedman, J., T. Hastie, and R. Tibshirani, 2000. "Additive logistic regression: a statistical view of boosting." *Annals of Statistics*. 28: 337-407.
11. Friedman, J., T. Hastie, and R. Tibshirani, 2003. *Elements of Statistical Learning*. Springer.
12. Gilson, R. & A. Schwartz, 2005. Understanding MACs: Moral Hazard in Acquisitions. *Journal of Law, Economics and Organization* 21 (2): 330-358.
13. Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M. and Dougherty, E.R. (2010). Small-sample precision of ROC-related estimates. *Bioinformatics* 26 (6): 822–830.
14. Hexion Specialty Chems., Inc. v. Huntsman Corp., 965 A.2d 715, 736 (Del. Ch. 2008).
15. NIXON PEABODY LLP, SEVENTH ANNUAL MAC SURVEY (2008), available at [http://www.nixonpeabody.com/publications\\_detail3.asp?ID=2488](http://www.nixonpeabody.com/publications_detail3.asp?ID=2488).
16. Macias, Antonio J., (2009). Risk Allocation and Flexibility in Acquisitions: The Economic Impact of Material-Adverse-Change (MACs) Clauses (Available at SSRN: <http://ssrn.com/abstract=1108792>).
17. Revlon, Inc. v. MacAndrews & Forbes Holdings, Inc. 506 A.2d 173, 184 (Del 1986).
18. Ripley, B.D., 1996. "Pattern Recognition and Neural Networks," Cambridge University Press.

19. Talley, E., 2009. On Uncertainty, Ambiguity, and Contractual Conditions.  
*Delaware Journal of Corporate Law*. 34(3), pp. 755-812.
20. Vapnik, V., 1995. *The Nature of Statistical Learning*. Wiley.

## Appendix 1: Screen shot from pilot study; Python-based interface.

*This screen shot displays a typical user interface as the program scours a sample acquisition agreement. In this screen shot, the program detects three species of MAC/MAE provisions, and fails to find seven others.*



```
Python Shell
File Edit Shell Debug Options Windows Help

Beginning automatic search for MAC relating to changes in trading price or trading volume of Company's stock...
Found an instance of MAC
    declared) or any escalation or worsening thereof;

    (viii) any change in the market price or trading volume of
    the shares of Class A Common Stock or IDs;

Beginning automatic search for MAC relating to changes in interest rates...
No instance of a MAC!
Beginning automatic search for MAC relating to changes in exchange rates...
No instance of a MAC!
Beginning automatic search for MAC relating to Acts of war or major hostilities...
Found an instance of MAC
    of the Company, holders of Notes of the Company or any other

    Person: (vii) any natural disaster or any acts of
    terrorism, sabotage, military action or war (whether or not

Beginning automatic search for MAC relating to Acts of Terrorism...
Found an instance of MAC
    of the Company, holders of Notes of the Company or any other

    Person: (vii) any natural disaster or any acts of
    terrorism, sabotage, military action or war (whether or not

Beginning automatic search for MAC relating to Acts of God...
No instance of a MAC!
Beginning automatic search for MAC relating to reduction of customers or decline in business...
No instance of a MAC!
Beginning automatic search for MAC relating to effects of announcing the transaction...
No instance of a MAC!
Beginning automatic search for MAC relating to changes caused by the taking of any action required in connection with the agreement...
No instance of a MAC!
Beginning automatic search for MAC relating to changes caused by the taking of any action required in connection with the agreement...
No instance of a MAC!
Input target company name:

Ln: 877 Col: 27
```

## Appendix II: Tables 1- 4

MAC/MAE Provision	Description	Freq.
MBOF	Affirmative MAC on the business, operations, financial condition, etc.	94.3%
MSelAbil	Affirmative MAC on Seller's ability to close the deal	48.8%
MExessLos	Affirmative MAC for losses over a specified threshold	2.4%
MPrspects	Affirmative MAC on prospects of the Company/Target	2.4%
MAssets	Affirmative MAC on the securities or other assets of target	21.1%
MReasExp	Affirmative MAC triggered if there is reasonable expectation of event to have a material adverse effect/change prospectively	12.2%
MDispEffect	Definition of materiality tied to "Disproportionate Effects"	73.2%
EChEcon	Exception for change in economy or business in general	82.1%
EChGen	Exception for change in general conditions of the specific industry	79.7%
EChSecM	Exception for change in securities markets	63.4%
EChPrVol	Exception for change in trading price or trading volume of Company's stock	52.0%
EChIntR	Exception for change in interests rates	18.7%
EChExch	Exception for change in foreign exchange rates	14.6%
EWar	Exception for acts of war or major hostilities	76.4%
ETerror	Exception for acts of terrorism	79.7%
EGod	Exception for acts of God	30.1%
ERedCust	Exception for reduction of customers or decline in business	29.3%
EAnnTran	Exception for effects of the announcement of the transaction	71.5%
EChAction	Exception for changes caused by the taking of any action required or permitted or in any way resulting from or arising in connection with the agreement	70.7%
EChGAAP	Exception for changes in GAAP	85.4%

**Table 1: MAC/MAE Terms and Relative Frequency in Data Set (n=123 deals)**



	<i>MBOF</i>	<i>MSelAbil</i>	<i>MExcessLos</i>	<i>MPrspects</i>	<i>MAssets</i>	<i>MReasExp</i>	<i>MDispEffct</i>	<i>ECbEcon</i>	<i>ECbGen</i>	<i>ECbSecM</i>	<i>ECbPrVol</i>	<i>ECbIntR</i>	<i>ECbExch</i>	<i>EWar</i>	<i>ETerror</i>	<i>EGod</i>	<i>ERedCust</i>	<i>EAnnTran</i>	<i>ECbAction</i>	<i>ECbGAAP</i>
<i>MBOF</i>	1																			
<i>MSelAbil</i>	0.099317	1																		
<i>MExcessLos</i>	0.038841	0.162019	1																	
<i>MPrspects</i>	0.038841	0.056578	-0.025	1																
<i>MAssets</i>	0.127181	0.132142	0.047227	-0.08186	1															
<i>MReasExp</i>	-0.12294	0.08365	0.102138	-0.058926	0.050463	1														
<i>MDispEffct</i>	0.326476	-0.10655	0.095743	-0.02321	0.043846	-0.11078	1													
<i>ECbEcon</i>	0.434773	0.200819	0.073794	0.0737939	0.085753	0.044274	0.291952	1												
<i>ECbGen</i>	0.39916	0.209977	0.07986	0.0798596	0.063563	-0.05873	0.286932	0.449594	1											
<i>ECbSecM</i>	0.104844	0.032121	-0.098746	-0.098746	0.103856	0.076743	0.18769	0.394241	0.245525	1										
<i>ECbPrVol</i>	0.185605	0.02541	0.046316	-0.059181	0.018795	0.009704	0.300104	0.188842	0.202526	0.047795	1									
<i>ECbIntR</i>	0.11781	-0.00916	-0.075829	-0.075829	-0.14615	0.076155	0.008035	0.11501	0.034965	0.19111	0.001357	1								
<i>ECbExch</i>	0.10171	-0.08193	-0.065465	-0.065465	-0.15802	0.056578	0.094967	0.133217	0.037642	0.17122	0.07524	0.686349	1							
<i>EWar</i>	0.359594	-0.07103	-0.160503	-0.03634	-0.04081	-0.02712	0.225628	0.24054	0.243004	0.174573	0.233455	0.119012	0.175784	1						
<i>ETerror</i>	0.224749	-0.03253	-0.18208	-0.05111	0.01408	0.003012	0.195737	0.238726	0.146531	0.245525	0.121647	0.138595	0.151962	0.766547	1					
<i>EGod</i>	0.161128	-0.28546	-0.10371	0.0112119	-0.03565	-0.1361	0.197126	0.121097	0.066974	0.019749	0.239448	0.140096	0.179828	0.364323	0.287238	1				
<i>ERedCust</i>	0.003762	-0.09155	-0.10171	0.0141263	-0.02669	0.033296	0.107216	0.067095	0.191691	0.228917	0.116896	0.103951	0.188656	0.062629	0.191691	0.006652	1			
<i>EAnnTran</i>	0.311741	0.038686	-0.017094	-0.017094	-0.11482	0.014773	0.228137	0.457941	0.397889	0.269162	0.115826	0.071386	0.159157	0.243995	0.174008	0.099342	0.049261	1		
<i>ECbAction</i>	0.150495	0.127297	-0.014126	-0.129962	0.114215	-0.0879	0.296073	0.259282	0.252338	0.216249	0.276537	-0.05812	0.013564	0.063656	0.163533	0.03231	0.295977	0.227954	1	
<i>ECbGAAP</i>	0.295442	0.173971	0.065465	-0.08365	-0.06733	0.08401	0.16461	0.34695	0.362477	0.306332	0.247057	0.139572	0.106349	0.257728	0.362477	0.171265	0.114673	0.350642	0.340321	1

**Table 2: Correlations Across MAC/MAE Provisions**

MAC/MAE Provision	<i>RE / Boolean Specification</i>		<i>LSA Specification</i>		<i>Hybrid Model</i>	
	Correctly Classified	ROC-AUC	Correctly Classified	ROC-AUC	Correctly Classified	ROC- AUC
MBOF	0.9431	0.5505 <i>0.1034</i>	0.9512	<b>0.81</b> <i>0.05</i>	0.9512	<b>0.8011</b> <i>0.0654</i>
MSelAbil	0.7886	<b>0.7865</b> <i>0.0366</i>	0.6260	<b>0.6646</b> <i>0.0492</i>	0.7886	<b>0.8608</b> <i>0.0336</i>
MExessLos	0.9754	0.56 <i>0.132</i>	0.9837	<b>0.8515</b> <i>0.1029</i>	0.9837	<b>0.8515</b> <i>0.1029</i>
Mprospects	0.9756	0.6625 <i>0.1667</i>	0.9756	<b>0.8472</b> <i>0.1132</i>	0.9837	<b>0.8444</b> <i>0.1288</i>
Massets	0.7886	<b>0.6887</b> <i>0.0489</i>	0.7886	0.5946 <i>0.0618</i>	0.7967	<b>0.7179</b> <i>0.0552</i>
MR easExp	0.8760	0.51 <i>0.1911</i>	0.8943	0.6596 <i>0.0815</i>	0.9008	<b>0.6664</b> <i>0.0805</i>
MDispEffct	0.7317	0.5965 <i>0.0506</i>	0.7561	<b>0.7215</b> <i>0.0593</i>	0.7886	<b>0.7182</b> <i>0.0603</i>
EChEcon	0.8211	0.5511 <i>0.0584</i>	0.8537	<b>0.7052</b> <i>0.0614</i>	0.8537	<b>0.7066</b> <i>0.0613</i>
EChGen	0.7967	0.572 <i>0.0557</i>	0.8049	<b>0.6865</b> <i>0.0619</i>	0.8211	<b>0.6918</b> <i>0.0607</i>
EChSecM	0.7317	<b>0.7462</b> <i>0.04</i>	0.6260	0.5942 <i>0.0526</i>	0.7642	<b>0.8132</b> <i>0.0407</i>
EChPrvol	0.7398	<b>0.7427</b> <i>0.0391</i>	0.6829	<b>0.7299</b> <i>0.0454</i>	0.7724	<b>0.8292</b> <i>0.0381</i>
EChIntr	0.8618	<b>0.6974</b> <i>0.0538</i>	0.8211	<b>0.7633</b> <i>0.0499</i>	0.8618	<b>0.8376</b> <i>0.0455</i>
EChExch	0.9024	<b>0.7817</b> <i>0.06</i>	0.8699	<b>0.8153</b> <i>0.0556</i>	0.9106	<b>0.882</b> <i>0.0562</i>
EWar	0.7642	<b>0.6812</b> <i>0.049</i>	0.7724	<b>0.7196</b> <i>0.0464</i>	0.7642	<b>0.7929</b> <i>0.0432</i>
Eterror	0.7967	<b>0.7461</b> <i>0.0414</i>	0.8130	<b>0.7276</b> <i>0.0513</i>	0.8049	<b>0.8137</b> <i>0.0435</i>
EGod	0.7724	<b>0.714</b> <i>0.0454</i>	0.7317	<b>0.7329</b> <i>0.0494</i>	0.7724	<b>0.7986</b> <i>0.0462</i>
ERedCust	0.8862	<b>0.8788</b> <i>0.0355</i>	0.7236	<b>0.7701</b> <i>0.0474</i>	0.8943	<b>0.9132</b> <i>0.0319</i>
EAnnTran	0.7154	<b>0.6153</b> <i>0.0495</i>	0.7236	<b>0.6722</b> <i>0.0557</i>	0.7317	<b>0.6946</b> <i>0.0552</i>
EChAction	0.7073	<b>0.6652</b> <i>0.0478</i>	0.7073	<b>0.69</b> <i>0.0584</i>	0.7073	<b>0.7663</b> <i>0.0475</i>

Table 3: Classification Rates and ROC-AUC for all data. Baseline Regression is as in Equation (2), which is estimated for each provision against three specifications: (a) RE Predictors only; (b) LSA Factors only; and (c) Hybrid RE and LSA. The classification protocol in the first column of each panel assigns term as present if the computed marginal probability evaluated at the estimated coefficients exceeds 0.5. The second column of each specification reports estimates of the area under the ROC. Standard errors are beneath, in italics.

	<i>LSA Specification</i>		<i>Hybrid Model</i>	
MAC/MAE Provision	Corr. Class	ROC-AUC	Corr. Class	ROC-AUC
MBOF	<b>0.913</b> <i>0.04744501</i>	0.61587 <i>0.2975</i>	<b>0.9284</b> <i>0.04117</i>	0.5327 <i>0.27813</i>
MSelAbil	0.54222 <i>0.08408</i>	0.5599 <i>0.0882</i>	<b>0.7421667</b> <i>0.070876</i>	<b>0.783655</b> <i>0.0784866</i>
Massets	<b>0.764333</b> <i>0.0675466</i>	0.6345 <i>0.1017</i>	<b>0.764</b> <i>0.068276</i>	0.58404 <i>0.10996</i>
MReasExp	<b>0.864033</b> <i>0.058078</i>	0.46393 <i>0.16247</i>	<b>0.8777333</b> <i>0.0555032</i>	0.49465 <i>0.190937</i>
MDispEffct	<b>0.7316</b> <i>0.067958</i>	0.6609 <i>0.10718</i>	<b>0.7357333</b> <i>0.07652223</i>	0.6320073 <i>0.1193881</i>
EChEcon	<b>0.8209</b> <i>0.05871753</i>	0.5367427 <i>0.1302458</i>	<b>0.8201667</b> <i>0.06222236</i>	0.6198363 <i>0.131313</i>
EChGen	<b>0.7823333</b> <i>0.06921313</i>	0.5047211 <i>0.1315335</i>	<b>0.7692</b> <i>0.06542236</i>	0.583982 <i>0.1247254</i>
EChSecM	0.5816 <i>0.07611767</i>	0.4832542 <i>0.09772</i>	<b>0.7038667</b> <i>0.07128283</i>	<b>0.7294512</b> <i>0.0914256</i>
EChPrvol	0.5678333 <i>0.07982975</i>	0.6334498 <i>0.087661</i>	<b>0.7281</b> <i>0.07246226</i>	<b>0.7751</b> <i>0.0852973</i>
EChIntr	<b>0.7874667</b> <i>0.06722013</i>	0.6621522 <i>0.132249</i>	<b>0.8251</b> <i>0.06456472</i>	0.7240246 <i>0.1269156</i>
EChExch	<b>0.8140667</b> <i>0.0616201</i>	0.5999294 <i>0.1569705</i>	<b>0.8785</b> <i>0.05493118</i>	<b>0.805348</b> <i>0.1496069</i>
EWar	<b>0.7290333</b> <i>0.06955984</i>	0.5679435 <i>0.100249</i>	<b>0.7242</b> <i>0.07423599</i>	<b>0.7217855</b> <i>0.0949628</i>
Eterror	<b>0.7728333</b> <i>0.06653955</i>	0.5361141 <i>0.1150063</i>	<b>0.7530333</b> <i>0.07099496</i>	<b>0.7372462</b> <i>0.1141708</i>
EGod	<b>0.6942333</b> <i>0.07621825</i>	0.5840398 <i>0.1060234</i>	<b>0.7248333</b> <i>0.07321809</i>	<b>0.727878</b> <i>0.097808</i>
ERedCust	<b>0.7230333</b> <i>0.08007837</i>	<b>0.7399407</b> <i>0.0928146</i>	<b>0.8596333</b> <i>0.05638499</i>	<b>0.8586069</b> <i>0.0849821</i>
EAnnTran	<b>0.6908667</b> <i>0.07399589</i>	0.3879988 <i>0.1039311</i>	<b>0.6821667</b> <i>0.07364824</i>	0.6091421 <i>0.110092</i>
EChAction	<b>0.7089</b> <i>0.0730488</i>	0.6196205 <i>0.105946</i>	<b>0.6671</b> <i>0.07533583</i>	<b>0.6711904</b> <i>0.1100768</i>

**Table 4: Monte-Carlo Simulation for out-of-sample prediction.** Baseline regression is as in Equation (2), estimated 1000 times on a "training" dataset containing a 75% sample (sampled randomly, for each iteration), and generating (simulated) out-of-sample predictions for the remaining 25%. For each term, the simulation explores two specifications: (a) LSA Factors only; and (b) Hybrid RE and LSA. The classification protocol in the first column of each panel assigns term as present if the computed marginal probability evaluated at the estimated coefficients exceeds 0.5. The second column of each specification reports estimates of the area under the ROC. Empirical standard errors appear beneath, in italics. (The MAC terms MExessLos and MProspects could not be simulated reliably because of their low-frequency representation in the data).